1

A DIALOGUE MODEL OF BELIEF

This paper offers a new model of belief by embedding the Peircean account of belief into a formal dialogue system that uses argumentation schemes for practical reasoning and abductive reasoning. A belief is characterized as a stable proposition that is derived abductively by one agent in a dialogue from the commitment set (including commitments derived from actions and goals) of another agent. On the model (to give a rough summary), a belief is defined as a proposition held by an agent that (1) is not easily changed (stable), (2) is a matter of degree (held more or less weakly or strongly), (3) guides the goals and actions of the agent, and (4) is habitually or tenaciously held in a manner that indicates a strong commitment to defend it. It is argued that the new model overcomes the pervasive conflict in artificial intelligence between the belief-desire-intention (BDI) model of reasoning and the commitment model.

This paper offers a definition of the notion of belief and a method for determining whether a proposition is a belief of an agent or not, based on evidence. The method is based on a formal dialogue system for argumentation that enables inferences to be drawn from commitments to beliefs using argumentation schemes (Walton, Reed and Macagno, 2008). Beliefs are defined as commitments that one really thinks hold true, that one is willing to defend against criticisms and objections, and that one uses as assumptions in deciding what to do and what is the case. Arguments are evaluated in a dialogue format using a set of critical questions matching each scheme. A defeasible scheme for inference to the best explanation is used to infer that an agent believes a proposition judging from its commitments and locutions in a dialogue. The approach offers a middle ground between the two leading artificial intelligence models that have been developed for programming intelligent agents. According to the commitment model, a commitment is a proposition that an agent has gone on record as accepting (Hamblin, 1970; 1971). In the BDI (belief-desire-intention) model (Bratman, 1987), intention and desire are viewed as the pro attitudes that drive goal-directed reasoning forward to a proposal to take action. The BDI model is based on the concept of an agent that carries out practical reasoning based on goals that represent its intentions and incoming perceptions that update its set of beliefs as it moves along (Wooldridge, 2002). In this paper, practical reasoning is used as an argumentation scheme for judging belief in a dialogue

As shown in a classic collection of essays (Engel, 2000), the proceedings of a conference on acceptance and belief, drawing the distinction between the two closely related notions has proved to be a hard problem to solve. The papers in the volume tried to grapple with this problem by drawing a precise distinction between acceptance and belief in various ways, but their attempts failed to reach agreement, primarily, it could be argued, because there was little or no basic agreement on how to define the notion of belief. As noted by Hinzen (2001 p. 282), writing about the dilemma about what an inquiry into belief should be about, a way out would be to simply define belief as a technical notion. However, there was no agreement in the collections of writings on belief he reviewed on how such a definition should be formulated.

Section 1 of the paper introduces the notion of practical reasoning as a form of inference, and as an argumentation scheme with a matching set of critical questions. Section 2 is taken up with the project of clarifying the notions of commitment and acceptance. Section 3 offers a pragmatic definition of the notion of belief, derived from the classic paper of Charles S. Peirce, ‘The Fixation of Belief’ (Peirce, 1877). In section 4, three formal dialogue models constructed in the Hamblin style are explained that are used to show how inferences from commitment to belief can be systematically drawn.

Section 5 shows how to use the dialogue systems and the new definition of belief to derive beliefs from commitments. Section 6 shows how, using this dialogue-based theory, we can build methods for determining in a given case whether a proposition is a belief of an agent or not. Section 7 shows how the argumentation scheme for abductive reasoning is the tool used to infer what an agent's beliefs are, based on eight types of positive evidence and four types of negative evidence. Section 8 presents a pair of defeasible argumentation schemes for argument from commitment to belief that can be used as additional inferential tools, and suggests some projects for future research.

1. Two Models of Agent Rationality

A rational agent of the kind described by Woodridge (2000), is an entity that has goals, some (though typically incomplete) knowledge of its circumstances, and the capability of acting in such a way as to seek to alter those circumstances. A rational agent also has the ability to perceive consequences of its actions, and to correct them if they move away from realizing current goals. An agent can be a person, or it can be an automated software program of the kind now commonly used on the Internet.[1] Some such systems are autonomous while others are tightly programmed in a deterministic way so that the actions lead to the goal by following a strict program. Autonomous agents have the capability of modifying their goals, and the capability of acting with other agents. The central kind of reasoning that agents use to navigate through their environments is called practical reasoning. Practical reasoning, in its basic form, consists of a chaining together of practical inferences of the sort represented by the simplest scheme for instrumental practical reasoning (Walton, 1996; Walton, Reed and Macagno, 2008, 323).

MAJOR PREMISE:	I have a goal *G*.
MINOR PREMISE:	Carrying out this action *A* is a means to realize *G*.
CONCLUSION:	Therefore, I ought (practically speaking) to carry out this action *A*.

This basic form of practical inference is very simple, yet we all recognize its importance as a kind of reasoning we use when deliberating on how to carry out a goal in circumstances where some action needs to be taken, and there can be reasons to accept or reject opposed proposals for action that are put forward. It has often been disputed in philosophy whether practical reasoning is purely instrumental or whether it needs to be based on values (Walton, 1990). There is also a more complex form of practical reasoning in which values are taken into account as well (Atkinson, Bench-Capon and McBurney, 2006). In this more complex model, a value is defined as a reason supporting a goal or reason counting against the adoption of a goal.

Practical reasoning, when deployed in a deliberation dialogue, gives a reason to accept a conclusion for a course of action tentatively, subject to exceptions or counter-arguments that may be advanced by the other side as new circumstances become known. The conclusion has a presumptive status, once reasons to support the proposed action are provided. However, such an argument is cast into doubt if any one of the following critical questions is asked (Walton, Reed and Macagno, 2008, 293).

[1] For this reason we feel free in this paper to refer to an agent as 'it' or 'she/he', depending on the context.

CQ_1: What other goals do I have that should be considered that might conflict with *G*?
CQ_2: What alternative actions to my bringing about *A* that would also bring about *G* should be considered?
CQ_3: Among bringing about *A* and these alternative actions, which is arguably the most efficient?
CQ_4: What grounds are there for arguing that it is practically possible for me to bring about *A*?
CQ_5: What consequences of my bringing about *A* should also be taken into account?

The presumptive status of the original argument is only restored if an appropriate answer to the critical question is given. Also, an instance of practical reasoning can be stronger or weaker as an argument, given the further arguments used to support it, or the opposing arguments or critical questions used to attack it.

Value-based practical reasoning is well explained by (Bench-Capon, 2003) and (Atkinson, Bench-Capon and McBurney, 2006). The following argumentation scheme for value-based practical reasoning is the one given in (Atkinson, Bench-Capon and McBurney, 2006, pp. 2-3).

In the current circumstances R
we should perform action A
to achieve New Circumstances S
which will realize some goal G
which will promote some value V.

According to this way of defining the scheme, values are seen as reasons that can support goals. This scheme also has a matching set of critical questions. The complete list of sixteen is given in (Atkinson *et al.*, 2006). Three of them are these ones.

- Will the action achieve the new circumstances?
- Will the action demote some other value?
- Is there another action which will promote the value?

The key factor to appreciate here is that practical reasoning needs to be evaluated in a dialogue context, like that of a deliberation where alternative proposal for actions are debated. Proposals are made on how to most effectively realize goals, and these proposals are critically questioned or attacked by arguing that an opposed proposal would do the job better. In such cases, parties need to share some goals and values, and also have access to shared data concerning the circumstances.

But what are goals? Are they intentions in the mind of the agent, or are they merely statements that the agent has expressed or formulated, either alone or as part of a group deliberation, and has pledged to carry out, or has publicly expressed commitment to? According to the commitment model, two agents (in the simplest case) interact with each other in a dialogue in which each contributes speech acts (van Eemeren and Grootendorst, 1992). Each has a commitment set, and as the one asks questions that the other answers, commitments are inserted into or retracted from each set, depending on the type of move (speech act) each speaker makes. The two participants in the dialogue take turns offering speech acts, like asking a question, making a proposal, or putting forward an argument. Each has a commitment set, and as each move is made,

commitments are inserted into or retracted from each set according to commitment rules, depending on the type of move each makes. A commitment is a proposition that an agent has gone on record as accepting (Hamblin, 1970; 1971). One type of speech act is the putting forward of an argument. On the commitment-based approach, practical reasoning is modeled in a dialogue format using an argumentation scheme, like the ones above, along with a set of critical questions matching the scheme. The reasoning is supported if the appropriate critical questions are properly answered, or it is otherwise undercut.

The word 'commitment' has acquired several different meanings in multi-agent systems (Maudet and Chaib-draa, 2002). Three of these can be noted here. First, a commitment is a persistent intention by an agent to undertake some action. Second, a commitment is a promise made by one agent to another to undertake some action. Third, a commitment is a dialogical obligation incurred by an agent in a dialogue to support some proposition when challenged in the dialogue. The third meaning is the one taken up as central in this paper, and it is the one modeled in dialogue systems like those of Hamblin cited above. The second meaning, of making a promise, is very important, but is beyond the scope of this paper. The first meaning takes us to consideration of the BDI model.

The second model, called the BDI (belief-desire-intention) model is based on the concept of an agent that carries out practical reasoning based on goals that represent its intentions and incoming perceptions that update its set of beliefs as it moves along (Wooldridge, 2002). On the BDI model (Bratman, 1987; Bratman, Israel and Pollack, 1988; Paglieri and Castelfranchi, 2005), an agent has a set of beliefs that are constantly being updated by sensory input from its environment, and a set of desires (wants) that are then evaluated (by desirability and achievability) to form intentions. On this model, the agent's goals in practical reasoning are represented by its intentions, persistent goals that are stable over time and are not easily given up. For example, on Bratman's (1987) version of the BDI model, forming an intention is described as part of adopting a plan that includes the agent's desires (wants) and beliefs.

The most distinctive difference between the commitment model and the BDI model is that desires and beliefs are private psychological notions internal to an agent, while commitments are statements externally and verbally accepted by an agent in a communicative context. The commitment model uses basic practical reasoning as an argumentation schemes, expressed in a dialogue format in which two or more agents deliberate to solve a problem. The BDI model uses intentions, beliefs, and desires representing a decision-maker's psychological states that are updated as the agent acquires new data from its perceptions. Despite these clear differences between the two models, many philosophers writing about practical reasoning often appear to blend them together. Searle (2001) ostensibly purports to advocate the BDI model of practical reasoning, but, like others, notably including Bratman (1987) and Levi (1997), very often uses the language of commitment to describe how agents engage in practical reasoning. In an influential paper, Bratman, Israel and Pollack (1988, p. 347) used this language to describe practical reasoning: "The fundamental observation of our approach is that a rational agent is committed to doing what she plans". Such wording here uses the language of commitment, although Bratman, Israel and Pollack portray themselves as advocating the BDI model.

Both models have been widely employed in artificial intelligence in multi-agent systems. Agent communication languages (ACL's) are used to enable agents to communicate with each other on the basis of conversational policies that are like dialogue rules. The following example of a BDI-based conversational policy states that one agent A can tell a second agent B something only if A believes it also and can establish that B does not believe it. But how is a programmer to implement this requirement? The programmer may be in no position to determine what A believes, or what A believes about what B believes. Hence Singh (1998, p. 40) and many other AI researchers in AI have moved to the commitment model.

Argumentation technology of the kind now widely being implemented in new AI systems for multi-agent computing (Wooldridge, 2002) has sometimes used the BDI model and sometimes used the commitment model (Dunne and Bench-Capon, 2006). Both are useful for evaluating practical reasoning under conditions of uncertainty in changing circumstances as leading to commitment where a conclusion is drawn on a basis of tentative acceptance (commitment) on a balance of considerations.

2. Acceptance and Commitment

The following four propositions seem fairly similar to each other, and each one might be interchanged with one or more of the others with little or no loss of meaning in everyday conversational argumentation. However, some fine distinctions between them can be drawn to bring out how each might be related to the other.

(A) Amanda is committed to the proposition 'Bruce was at the bank yesterday'.
(B) Amanda accepts the proposition 'Bruce was at the bank yesterday'.
(C) Amanda is of the opinion that Bruce was at the bank yesterday.
(D) Amanda is convinced that Bruce was at the bank yesterday.
(E) Amanda believes that Bruce was at the bank yesterday.

Proposition A could be taken to mean that Amanda has gone on record as stating the proposition that Bruce was at the bank yesterday. It could be taken to mean that Amanda has stated that Bruce was at the bank yesterday, or has voiced other opinions that imply this. For example, she may have testified under oath in court that Bruce was at the bank yesterday, and so she is very definitely committed to that proposition now. Proposition B seems similar to proposition A, except that it seems more to suggest some explicit act of acceptance of a proposition, as contrasted with some underlying commitment to it that may be binding. Proposition C could be taken to mean that she is not only of the opinion that Bruce was that the bank yesterday but she has strong enough reasons to justify having this opinion. Proposition D appears to imply all of the propositions above it. D suggests that Amanda is fairly firmly committed to the proposition that Bruce was at the bank yesterday, and has some reasons to think it is true. Proposition E would appear to imply that she is convinced that Bruce was at the bank yesterday, and is strongly convinced as to resist admitting that she is mistaken. It is interesting to compare brief and inconclusive remarks about the natural language meanings of the four terms in question to the way they are used as terms of art in fields like artificial intelligence and epistemology.

An example due to Goldstein (Engel, 2000, p. 65) was put forward to show not only the difference between belief and acceptance, but that there can be a conflict between the two notions in a given case: “Typically, a person who plays the stock market believes, deep down, that the only person likely to win is the stockbroker, yet, driven by greed, accepts the stockbroker’s forecast that he, the punter, will make stupendous financial gains”. Cohen’s pioneering analysis of the distinction between belief and acceptance (1992) also showed that acceptance of proposition *A* is, in principle, compatible with the belief that not-*A*. Examples of this sort show not only the possibility of drawing a distinction between belief and acceptance, but also the importance of drawing it. As Engel showed in his introduction to the conference proceedings on belief and acceptance (2000), the distinction, if it could be clarified, would have philosophical power as applied to solving philosophical problems like Moore’s paradox. Moore’s paradox is the puzzle that the expression ‘*P* but I don’t believe that *P*’, seems to be an absurd assertion, even though it is not logically inconsistent. An example is the statement, ‘It is raining, but I do not believe it is raining’. It is different from the statement, ‘It is raining, but I am not committed to the proposition that it is raining’. These statements, as will be shown below, need to be treated in different ways as “paradoxical”.

Is acceptance the same as commitment? In other words, is the statement ‘Bob accepts *A*’ always equivalent to the statement ‘Bob is committed to *A*’? An agent’s commitment is defined as applying to any statement that he has gone on record in a dialogue as committing himself to, in virtue of what he said (or did, in the case of action commitments) in the past (Hamblin 1970; Walton and Krabbe, 1995). Thus commitment is a kind of dialogue notion that underlies speech acts. Part of defining any speech act is determined by commitment conditions that apply to the speaker and hearer when this speech act is uttered in a dialogue. This same analysis also appears to partly apply to acceptance. Acceptance, on the one hand, is a kind of speech act. When I say ‘I accept *A*’, my uttering this speech act can be taken to imply I am now committed to *A*, possibly subject to later retraction. On the other hand, acceptance is a more general notion that underlies all speech acts and sets requirements on how each speech act is to be defined as a form of dialogue move. In this second sense, it is equivalent to commitment.

Hamblin (1970) looked around for a model of rational argument that could be used as a basis for studying informal fallacies. One might propose epistemological models based on knowledge and belief. For example a strong rational argument might be one in which the premises are known to be true and logically imply the conclusion. Or a less strong rational argument might be one in which the premises are believed to be true and logically imply the conclusion. Hamblin rejected such models (1970, chapter 7) because of difficulties in formalizing a notion of rational argument based on modal operators of knowledge and belief (238). He was led, for these reasons, to take the alternative route of choosing acceptance, rather than knowledge or belief, as his basic notion on which to build his theory of rational argument. He formulated the following four dialogical criteria of the justifiability of an argument (1970, p. 245).

(D1) The premises must be accepted.

One might ask here, ‘accepted by whom?’ The answer Hamblin gave is that they should be ‘accepted by X’, where X is the name for some person or group of persons that remains in contact throughout the argument.

(D2) The passage from premises to the conclusion must be of an accepted kind.

This requirement has been expanded by recent developments in argumentation theory, so that it would include not only deductive and inductive forms of argument as warrants, but also argumentation schemes of the presumptive sort.

(D3) Unstated premises must be of a kind that are accepted as omissible.

This criterion requires some analysis of the notion of an enthymeme, or argument containing premises (or a conclusion, possibly) not stated in the given text of discourse. Generally, an argument, in the sense of such a set of criteria, is taken as something given, or expressed in a text of discourse. But some premises (or the conclusion) may not be explicitly expressed in the discourse, and may have to be inferred as assumptions on which the argument depends.

(D4) The conclusion must be such that, in the absence of the argument, it would not be accepted.

This condition expresses the notion that an argument, as evaluated by the four criteria, is something that gives a reason to accept the conclusion. For the argument to be of worth, it must boost up the evidential weight of the conclusion as a proposition that is to be accepted. Further elaboration of (D4) thus involves the notion of being able to judge the value of the worth of a proposition or argument for acceptance. But Hamblin did not attempt to present any further elaboration of such notions as part of his four dialectical criteria of the worth of an argument.

A key passage in Hamblin’s analysis (1970, 246) occurs where he asked why he used the word ‘accepted’ in his primary formulation rather than the word ‘believed’. The reason he offered is that the term ‘believed’ is too much of a psychological word, “conjuring up pictures of mental states”. In contrast he saw the term ‘acceptance’ as something that relates to putting on a “linguistic performance”. Nowadays that would be called a speech act. His exact wording (1970, 257) is interesting:

> A speaker who is obliged to maintain consistency needs to keep a store of statements representing his previous commitments, and require of each new statement he makes that it may be added without inconsistency to this store. The store represents a kind of *persona* of beliefs; it need not correspond with his real beliefs, but it will operate, in general, approximately as if it did. We shall find that we need to make frequent reference to the existence, or possibility, of stores of this kind. We shall call them *commitment-stores:* they keep a running tally of a person’s commitments.

Commitments, on Hamblin’s view, are public and social. If you make an assertion of a statement *A* in a way that indicates you are committed to it, and there is a public record of your speech act of asserting *A* in this manner, then that is evidence you are committed to *A*. For example, if you confess to a murder under police questioning, and the interview

was videotaped, then the videotape provides evidence that you are committed to the statement that you murdered the victim. In law, the videotape itself is called evidence, and when it is shown in court, it provides evidence for the accusation that you are guilty of the crime as alleged. Thus once you have committed yourself to a statement, say by asserting it in public so that your assertion can be recorded or be put "on record", then that is evidence of your commitment to it. Thus commitment is inherently a social notion that has to do with public dialogues in which two parties or more engage in public conversations. Commitment is basically public. Your commitments are inferred from what you have gone on record as saying in some context of dialogue.

Belief, although it can sometimes be public, as when we talk about commonly held beliefs, is a more private matter. If belief is an internal psychological matter of what an individual really thinks is true or false, the privacy of belief makes it more difficult to judge what an individual believes. People often lie, or conceal their real beliefs. And there is good reason to think that people often don't know what their own beliefs are. If Freud was right, we also have unconscious beliefs that may be quite different from what we profess to be our beliefs. Belief is deeply internal and psychological, and public commitment to a proposition is not necessarily an indication of belief. But perhaps there is a way to infer belief from commitment.

Two forms of an argumentation scheme called argument from commitment are presented in the compendium of schemes in (Walton, Reed and Macagno, 2008, 335). The first form is the simpler version.

COMMITMENT EVIDENCE PREMISE:	In this case it was shown that *a* is committed to proposition *A*, according to the evidence of what he said or did.
LINKAGE OF COMMITMENTS PREMISE:	Generally when an arguer is committed to *A*, it can be inferred that he is also committed to *B*.
CONCLUSION:	In this case, *a* is committed to *B*.

The second form is more complex, and refers to a dialogue (see section 4 below).

MAJOR PREMISE:	If arguer *a* has committed herself to proposition *A*, at some point in a dialogue, then it may be inferred that she is also committed to proposition *B*, should the question of whether *B* is true become an issue later in the dialogue.
MINOR PREMISE:	Arguer *a* has committed herself to proposition *A* at some point in a dialogue.
CONCLUSION:	At some later point in the dialogue, where the issue of *B* arises, arguer *a* may be said to be committed to proposition *B*.

Two critical questions match each scheme (Walton, Reed and Macagno, 2008, 335).

CQ_1:	What evidence in the case supports the claim that *a* is committed to *A*, and does it include contrary evidence, indicating that *a* might not be committed to *A*?
CQ_2:	Is there room for questioning whether there is an exception in this case to the general rule that commitment to *A* implies commitment to *B*?

The problem is how the bridge between commitment and belief can be crossed. That is, how can one draw a rational inference from a person's commitment to a statement to the conclusion that he believes that this statement is true? The inference is surely a hazardous one in many instances. A participant in a discussion will often make or incur commitment to some proposition for the sake of argument without really believing that proposition, or even being in a position to know for sure whether it is true or not. However, argument from commitment is a defeasible argumentation scheme, and this aspect of it might be quite favorable for using it to argue from commitment to belief.

3. A Peircean Definition of Belief

Why do we need a notion of belief, as opposed to commitment? There are basically three reasons. The first is that western philosophy since the time of Plato has taken the notions of knowledge and belief to be closely connected, and has therefore taken belief to be a fundamental notion of epistemology. Epistemology could even be defined as the philosophy study of knowledge and belief. One could even argue that knowledge implies belief by considering the following statement, a variant on Moore's paradox: I know that the sky is blue but I don't believe it. This statement is arguably self- contradictory, and if so, one could argue that knowledge implies belief. Indeed, philosophy has traditionally defined knowledge as justified true belief, even though counterexamples have shown this definition to be a highly problematic, if not untenable. In short, the one reason for thinking that we need a notion of belief, as opposed to commitment, is that traditional epistemology, which has strongly influenced models of rational thinking in artificial intelligence, has held belief to be a fundamental notion closely connected to the fundamental notion of knowledge. Mainstream psychology has also accepted this tradition by taking belief to be the most basic type of mental representation, and by assuming it is a building block of conscious thought.

The second reason is that it can be plausibly argued that we need the notion of belief in order to connect abstract models of rational thinking to real cases of argumentation (Godden, 2009). For example, in persuasion dialogue, the proponent's goal is to persuade the respondent to come to accept some proposition that he didn't accept before. Persuasion, on this model, is defined as a change in the respondent's commitments set over a sequence of moves in the dialogue. What has to happen for persuasion to occur is that the respondent was not committed to the proposition *A* before the proponent made her persuasion attempt, but then he became committed to *A* after the persuasion attempt. This structure defines successful persuasion in the formal model. However, in real life, it might be argued that you haven't really persuaded somebody of *A* successfully unless you have brought about that he believes *A*. On this approach, rhetorical persuasion, to be successful, must achieve not just a change in the respondent's commitment to the target proposition, but must really change the respondent's belief with respect to that proposition. In order to be of any assistance to fields like psychology and rhetoric, that take belief change to be fundamental to understanding thinking and persuading, we need the notion of belief.

Despite widespread acceptance by philosophers that the notion of belief is fundamental to epistemology, there is also been disagreement and controversy (Rudder

Baker, 1989). Some philosophers hold that beliefs can be modeled as types of sentences, and that the common sense understanding of belief is correct. Others have argued that the common sense understanding of belief is wrong, and that there is no coherent mental representation of this common sense notion. Some have even argued that the common sense concept of belief is obsolete, much like the notion of phlogiston in the discredited scientific theory of combustion (Churchland, 1981; Stitch, 1983). If these skeptical views about belief are right, there are significant consequences for science, including not only artificial intelligence generally, but also fields like neuroscience, where belief is taken to be a central concept.

The third reason has to do with negative concepts like insincerity, self-deception and lying, all of which appear to require some notion of belief. For example, the speech act of telling a lie could be defined as putting forward a statement as true when one believes (or even knows) that it is false. These concepts are fundamentally important not only in ethics, but also important in law in the process of examination in trials (including cross-examination), as well as in witness testimony and the crime of perjury. It is one thing to commit yourself in a dialogue to a proposition that that you are not really committed to, as judged by your prior commitments in the dialogue. There might be many reasons to explain such an inconsistency of commitments. Perhaps you just forgot, or you can somehow explain the inconsistency. Maybe you just changed your mind, as some new evidence came into the dialogue. But lying is a different thing. To lie, you have to really believe that the statement you made is false. In short, negative notions of a significant kind, like lying, self-deception, and so forth, cannot be fully understood only through applying the notion of commitment, but also require reference to belief. Lying is also closely related to notions like lying by omission, equivocation, deception, and using ambiguity in argumentation, and these notions are in turn related to the study of informal fallacies.

The best place to start in working out an analysis of belief is to start with some central characteristics of the notion that are widely accepted. Engel (1998), following the analysis of Bratman (1993), based on the prior analysis of (Kaplan, 1981, 1981a), identified five leading characteristics of belief (summarized by Tuomela, 2000, p. 122).

(1) Beliefs are involuntary, and are not normally subject to direct voluntary control.
(2) Beliefs aim at truth.
(3) Beliefs are evidence-related in that they are shaped by evidence for what is believed.
(4) Beliefs are subject to an ideal of integration or agglomeration.
(5) Beliefs come in degrees.

The problem is how to use these characteristics to develop a definition of the concept of belief that would be useful in artificial intelligence to assist inquiry into topics like how to draw a distinction between acceptance and belief. Consider the first characteristic: beliefs are not normally subject to voluntary control. The problem here might seem to be that in order to define ‘belief’, we first have to define the notion of voluntary control. However, one can define what a certain kind of thing is even if one does not have a definition of one of its properties. For example, one can define what oxygen is (an element whose nucleus contains eight protons) even if one cannot define its property of causing oxidation.

The second characteristic shows that it makes sense to speak of beliefs as true or false. There is a contrast with commitments here. The proposition one is committed to can be true or false, but commitments are accepted or not. However, to make this characteristic useful for our analysis of belief to follow, we have to define how belief is related to truth, or aims at truth. Beliefs should aim at truth, but truth can be hard to find, and therefore the characteristic that beliefs are evidence-related is very important. Note also that if we accept the proposition that beliefs come in degrees (characteristic 5), modeling belief as an either-or property, as done in modal logics of belief, is inappropriate. Degrees of belief are best modeled by a quantitative apparatus, for example by using real numbers between 0 and 1 to represent a person's degree of belief that a proposition is true.

There is, however, a way in which these five characteristics could be used to orient us to a way of defining the notion of belief in a pragmatic manner that could be useful in artificial intelligence, cognitive science, and other fields of investigation that need to rely on some clear notion of belief. In accord with characteristics 2 and 3, it can be argued that belief should be defined not only as a propositional attitude, but as an evidence-related notion that tracks the way beliefs are fixed and modified during an investigation in which evidence is being collected in order to prove some ultimate proposition that is in doubt. To say that a proposition is in doubt means that it is unstable in a certain sense: there is insufficient evidence to prove or disprove it, as far as the investigation has proceeded to this point. Therefore the investigation should be continued and further evidence should be collected in order to overcome this instability and lead to a stable fixation of belief. Looking at belief this way, it needs to be defined not only as a propositional attitude but in a pragmatic context of investigation prompted by doubt or instability, but with the aim of moving towards finding the truth of the matter being investigated. The problem of defining belief can now be recast as a pragmatic problem of the fixation of belief in a single agent or group of agents[2] conducting an investigation in which evidence is being collected and evaluated.

Charles S. Peirce, in his classic essay on the fixation of belief (1877), offered some clues on what he took belief to be. Peirce's essay was about what he called the "fixation' of belief, meaning how belief is "fixed", or lodged in place in an agent, and how it is removed or changed. He contrasted the method of authority with the method of science as ways belief is fixed and changed. Peirce did not attempt to define belief, but he did offer a description of what belief is, and how it works, by contrasting it with a characterization of doubt, and how it works. A quotation from this account shows that his description of how belief works presents several of its leading characteristics.

> We generally know when we wish to ask a question and when we wish to pronounce a judgment, for there is a dissimilarity between the sensation of doubting and that of believing. Our beliefs guide our desires and shape our actions. The Assassins, or followers of the Old Man of the Mountain, used to rush into death at his least command, because they believed that obedience to him would insure everlasting felicity. Had they doubted this, they would not have acted as they did. So it is with every belief, according to its degree. The feeling of believing is a more or less sure indication of there being established in our nature some habit which will determine our actions. Doubt never has such an effect.

[2] We have not taken up the notion a single agent trying to determine whether he believes something or not (Meijers, 2002), confining the discussion to brief consideration of how a pair of agents engage in dialogue. Still, the possibility that one can deliberate with oneself by examining the pros and cons of an issue could fit the dialogue model.

Nor must we overlook a third point of difference. Doubt is an uneasy and dissatisfied state from which we struggle to free ourselves and pass into the state of belief; while the latter is a calm and satisfactory state which we do not wish to avoid, or to change to a belief in anything else. On the contrary, we cling tenaciously, not merely to believing, but to believing just what we do believe. Thus, both doubt and belief have positive effects upon us, though very different ones. Belief does not make us act at once, but puts us into such a condition that we shall behave in some certain way, when the occasion arises.

The Assassins were members of a militant religious sect, thought to be active in the 8th to 14th centuries, who sent out members to carry out politically motivated assassinations. The use of this example suggests that Peirce associated belief with religious or group beliefs of similar kinds that can be dogmatic in nature, or based on the authority of a spiritual leader. Thus by belief, he does not necessarily mean rational belief. Peirce takes belief to be manifested in habits of inferring, especially inferring what one is to do (and doing it). The belief that obedience to the commands of the Old Man of the Mountain will bring everlasting felicity is manifested in the habit of inferring from the fact that the Old Man of the Mountain has commanded something that one is to do it right away.

Careful reading of Peirce's description and accompanying remarks suggests that belief is taken to possess ten defining properties.

I Stable State

(1) It is opposed to doubt, an uneasy and dissatisfied state.
(2) It is a settled state, a "calm and satisfactory state".

II Not Easily Changed

(3) It is a state we do not wish to change.
(4) It is something we cling tenaciously to.
(5) We cling to believing what we believe.
(6) It can be firmly fixed, as with fanatical believers.
(7) It is an indication of a habit.

III Matter of Degree

(8) It is a matter of degree.

IV Related to Future Actions

(9) It puts us into a condition so we act in a certain way in the future.
(10) It both guides our desires and shapes our actions.

Point (7) needs to amplified by clarifying what sort of habit belief indicates. Judging from the Man of the Mountain example quoted above, Peirce refers to habits of inferring, for example the habit of drawing a conclusion to act based on the command of the Man of the Mountain. However, throughout his paper, he writes of habits of mind that that lead us to infer factual conclusions, in addition to habits of mind that lead us to act in certain ways. It might be suggested that what Peirce refers to are rather like

argumentation schemes, defeasible and habitual forms of reasoning that are used to draw a conclusion to act, or to accept the conclusion, based on standardized premises. For example, the conclusion to do what the Man of the Mountain says is presumably based on some kind of argument from authority. However, such habits are not always based on reasonable forms of argument. They might also be based on fallacious or very weak forms of inference that are nevertheless acted on by an agent.

Based on these observations, we could set out a Peircean definition of belief by, first of all, stating the implicit assumptions that belief is something that is a property of an agent, and can be said to be held by that agent, and then by adding the properties to this base. An agent can be defined as an entity that can carry out actions based on data that it can receive and that can perceive (to a limited extent) the consequences of its actions. A practical reasoning agent is one that has information about its environment, including ways of carrying out actions, and uses this information in order to attempt to bring about some ultimate goal. An investigating agent is one that has access to evidence, and that collects and evaluates this evidence in order to try to prove or disprove some ultimate proposition. The proposed definition reads as follows:

Belief is (1) a stable state of an agent (2) that is not easily changed, (3) is a matter of degree (held more or less weakly or strongly), (4) that guides the desires and actions of that agent in goal-directed reasoning in deliberation, or the evidential thinking of that agent in an investigation aimed at removing doubt, and (5) the content of a belief is a proposition to which the agent would ordinarily assent to if asked, unless the agent is unaware having the belief or wishes to conceal the fact of believing the proposition from the questioner.

This definition is a pragmatic one in that it postulates two different kinds of frameworks in which agents can be said to have beliefs.

Now we have offered a definition of the notion of belief that we can work with, the next problem is to find a dialogue framework that could be used to model the operation of one party using the commitments of another to derive conclusions about what the first party believes from his/her commitments.

4. Dialogue Systems for Explicit and Implicit Commitment

The solution to the problem is to adapt a formal dialogue system that can model these notions in a limited but very useful way, and define commitment and belief in the model. The dialogue system CB[3] (Walton, 1984), a formal system devised after the fashion of the formal dialogue systems of Hamblin's earlier systems, has just four locution rules, five commitment rules and three dialogue rules (Walton 1984, pp133-135).

Locution Rules

(i) Statements: Statement-letters, S, T, U, . . . , are permissible locutions, and truth-functional compounds of statement-letters.

[3] The names of the dialogue systems started with the letters A, AA, AB, B, BA, and so forth. CB was the third of the C systems. The idea was to start with the simplest systems and work up to more complex ones.

(ii) Withdrawals: ‘No commitment S’ is the locution for withdrawal (retraction) of a statement.
(iii) Questions: The question ‘S?’ asks ‘Is it the case that S is true?’
(iv) Challenges: The challenge ‘Why S?’ requests some statement that can serve as a basis in proof for S.

Commitment Rules

(i) After a participant makes a statement, S, it is included in his commitment store.
(ii) After the withdrawal of S, the statement S is deleted from the speaker’s commitment store.
(iii) ‘Why S?’ places S in the hearer’s commitment store unless it is already there or unless the hearer immediately retracts his commitment to S.
(iv) Every statement that is shown by the speaker to be an immediate consequence of statements that are commitments of the hearer then becomes a commitment of the hearer’s and is included in his commitment-store.
(v) No commitment may be withdrawn by the hearer that is shown by the speaker to be an immediate consequence of statements that are previous commitments of the hearer.

Dialogue Rules

(Rl) Each participant takes his turn to move by advancing one locution at each turn. A no-commitment locution, however, may accompany a why-locution as one turn.
(R2) A question ‘S?’ must be followed by (i) a statement S, (ii) a statement ‘Not-S’, or (iii) ‘No commitment S’.
(R3) ‘Why S?’ must be followed by (i) ‘No commitment S’ or (ii) some statement T, where S is a consequence of T.

Strategic Rules

(i) Both participants agree in advance that the dialogue will terminate after some finite number of moves.
(ii) The first participant to show that his own thesis is an immediate consequence of a set of commitments of the other participant wins the dialogue.
(iii) If nobody wins as in (ii) by the agreed termination point, the dialogue is declared a draw.

Whatever rules of inference we adopt, the definitions of immediate consequence and consequence from (Walton, 1984, 132-133) need to hold.

Definition of ‘immediate consequence’: A statement T is an immediate consequence of a set of statements S0, S1,..., Sn if and only if ‘S0, S1,..., Sn therefore T’ is a substitution-instance of some rule of the game.

Definition of ‘consequence’: A statement T is a consequence of a set of statements S0, S1,..., Sn if and only if T is derived by a finite number of immediate consequence steps from immediate consequences of S0, S1, ... Sn.

It will be shown below that to use such dialogue systems to model the notion of a participant, the simple system CB is not enough, and we have to go on to add an additional feature by drawing a distinction between explicit and implicit commitments.

The problem with the principle that belief always implies commitment is that commitment to a proposition, on Hamblin’s view, means that a participant in a dialogue has gone on record as committing himself to the proposition. In other words, on Hamblin’s view, commitment always refers to the explicit kind of commitment that a participant in dialogue has made by making a verbal statement of acceptance. However, this is a rather narrow notion of commitment, for several reasons. One is that there may be reasons why a participant in the dialogue can’t go on record as stating in public that he believes a particular proposition. For example, making such a statement may be illegal or politically incorrect, or indeed it may be a statement that the participant is committed not to asserting in public, as it would reveal confidential information he has sworn on oath not to reveal. Another reason is that a participant may not be aware of all his beliefs. He may be deeply committed to some proposition, but not even realize that himself, having not given the matter much thought. But under examination by another party in circumstances that force him to make some difficult choices, he may come to realize that he is committed to this proposition, even though he was unaware of the previously. In other words here we can draw a distinction between explicit commitments that he has gone on record as accepting in a dialogue and implicit commitments, representing propositions that he is really committed to, even though he has never stated them or articulated them in public.

If we think of the concept of commitment in this broader meaning so that it includes implicit commitments as well as explicit commitments that a participant in dialogue has gone on record as stating or accepting, it is possible to define belief as a species of commitment. So defined, belief is a commitment that a participant strongly holds in a manner indicating that he not only accepts it, but includes it among the things that he regards as having some standing or veridicality, and is willing to defend it as holding, even in the face of objections and criticisms. Commitments include all kinds of propositions that one may have gone on record as accepting for the sake of argument, that one has agreed to because one has no particular reason not to. Beliefs, on the other hand, have to meet a higher standard. They represent propositions that one really thinks hold true, even though one may be persuaded by counter-evidence that is sufficiently strong. Beliefs can be retracted, but only if there is enough evidence against them to outweigh the reasons one has for accepting them.

To model the notion of belief, we need to go beyond the systems devised by Hamblin that modeled a commitment set in a dialogue as a set of public statements, on display in view of all the participants. In CBV, each commitment set divided into two subsets, one consisting of the explicit commitments a party has gone on record as asserting, the other consisting of a set that neither party can see unless some move in the dialogue reveals one of them. In (Walton, 1984) the implicit commitment were called dark-side

commitments and the explicit commitments were called light-side commitments.[4] A participant is only explicitly committed to statement S if he has gone on record in a previous move of the dialogue as asserting S. But what if he asserts two other statements T and U that logically imply S? In some sense, he is then implicitly committed to S. Or what if he puts forward an argument that would be valid only if the relatively uncontroversial premise S is added to it? Once again, we might be inclined to say that he has implicitly committed himself to S. As a CBV persuasion dialogue continues, more statements come over from the implicit (dark) side to the explicit (light) side in the commitments sets of a participant. Any time he denies commitment to S, but S is really an implicit commitment of his, S will immediately appear among his explicit commitments. Such a dialogue tends to reveal the implicit commitments of an arguer.

CBV has one additional commitment rule, Rule vi, in addition to the rules of CB.

Commitment Rule (vi): If a participant states 'No commitment S' and S is on the implicit side of his commitment store, then S is immediately transferred to the explicit side of his commitment store.

The question now is how we can apply CBV to determine in a given case where a dialogue is underway whether or not some proposition is an implicit commitment of a participant or not. Basically, the method is to question the participant whether or not he accepts this proposition in light of other propositions he has committed himself to in the past sequence of the dialogue. As a CBV persuasion dialogue continues, more statements come over from the implicit (dark) side to the explicit (light) side in the commitments sets of a participant. Any time he denies commitment to S, but S is really an implicit commitment of his, S will immediately appear among his explicit commitments. Such a dialogue tends to reveal the implicit commitments of an arguer.

Reed and Walton (2007) constructed an extension of CB called ASD that incorporates argumentation schemes[5]. The move 'pose(C)' allows for the asking of a critical question selected from those matching the scheme. In CB dialogues, the 'S … Why S? … T' sequence of argumentation (where T is offered as a reason given to support S) is a common pattern. The characteristic feature of ASD is that it has a new dialogue rule R4 that applies after this sequence.

(R4) After a statement T has been offered in response to a challenge locution, Why S?, then if (S, T) is a substitution instance A of some argumentation scheme of the dialogue, the locution pose(C) is a legal move, where C is a critical question of scheme A appropriately instantiated.

What is needed to analyze argumentation about beliefs is a system ASDV that has rule R4 added to CBV. In this system one participant can use argumentation schemes to draw plausible inferences about the other party's beliefs. He can then ask the other party if he agrees with the belief attributed to her, or he can use arguments based on schemes to question whether the other party has a particular belief or not.

[4] The V stands for veiled or dark-side (implicit) commitments.
[5] ASD stands for argumentation scheme dialogue.

CB, CBV, ASD and ASDV are meant be very simple, basic dialogue systems to which more specialized rules can be added to model specific types of dialogue. Such specific dialogue types tend to be much more complex, and require many more rules (Walton and Krabbe, 1995). One type of dialogue in particular that is centrally important as a method for determining the beliefs of an agent is examination dialogue (Dunne et al., 2005; Walton, 2006). There is no space here to go into the details of such complex dialogue systems, however. All that is needed are some basic systems that explain commitment.

5. Deriving Belief from Commitment

We now come back to the original problem of trying to show how belief derives from commitment in dialogues, and can be seen to represent a special type of commitment. In CBV the distinction was drawn between a commitment that may have been incurred merely for the sake of argument, and may easily be retracted, and a commitment incurred by reason of making a strongly voiced claim of the kind one is obliged to defend. The second type of commitment is more firmly fixed, and one may be more reluctant to give up such a commitment even in the face of reasons for giving up. It could be hypothesized that one difference between the two concerns how each type of commitment represents a state that is stable and firmly fixed, and that therefore the second kind of commitment, because it is more stable and firmly fixed, is more closely related to the notion of belief .

It is generally accepted that explicit commitment does not necessarily imply belief, as the comments of Hamblin (1971) make clear. I may go on record as saying something, and thereby commit myself to that proposition, even though I do not believe it. For example, I may be lying or deceiving myself. It would appear that belief does not imply explicit commitment, as I may believe something without ever mentioning it. Nor does belief imply commitment, since someone can believe something without ever indicating the existence of that belief.

If the foregoing remarks are correct, it would appear that there is no logical relationship between commitment and belief. This appearance is misleading, however, once we take implicit commitment into account. What the foregoing remarks show is that there is no connection of deductive logical implication between explicit commitment and belief and vice versa. The possibility remains that there may be a connection in a system like CBV that allows for both explicit and implicit commitments. This connection can be explained as follows. Suppose that you believe a particular proposition *A*, and *A* is not in your commitment set, nor is there any subset of propositions within your commitment set that logically implies *A*. Still, it may be the case that you believe that proposition *A* is true. What then is the link between your commitment set and your belief that proposition *A* is true? The link is that I can engage in an examination dialogue with you about proposition *A*, and about other factual propositions related to *A*, and judge from the commitments I can extract from you in this dialogue whether you believe proposition *A* or not. I can even ask you directly whether you believe *A* or not. Even if you claim not to believe *A*, I can ask you whether other propositions you have shown yourself to be committed to in the dialogue imply belief in *A*. So we can say that although there is no link of deductive logical implication between belief and explicit commitment, there can be defeasible links between sets of one's commitments, both implicit and explicit. An agent can be questioned in CBV as to whether his commitments imply a certain belief,

and it may even be put to the agent that some set of his explicit and implicit commitments reasonably implies, on the basis of defeasible argumentation schemes, that he believes a particular proposition. The agent may deny that he believes in this proposition, but if what he said has on more than one occasion defeasibly implied belief in this proposition, the case can be made that there is evidence for such a belief, based on dialogue with the agent in which he was questioned about his commitments. Take Moore's paradox as an example. Suppose that agent says '*P* but I don't believe that *P*'. To confront the apparent inconsistency, we have to examine the agent's commitment set, and see whether there is some proposition *P* somewhere, and if there is evidence that it is a stable commitment that the agent has often strongly defended. Now we have to ask him why he claims no longer to believe *P*, given that there is evidence he did before. The agent's assertion '*P* but I don't believe that *P*' needs to be handled differently if there is no such evidence. He needs to be asked why he is now saying something that he doesn't believe. Maybe, for example, he is just conceding it for the sake of argument, even though his grounds for accepting it are not, in his view, strong enough to justify his accepting it. The point is that different responses are required in CBV, depending on the circumstances, and the apparently paradoxical statement can be dealt with in appropriate ways.

Now we can understand why stability and reluctance to change a commitment are characteristics of a belief. The reason is that stability and reluctance to give up a proposition indicate a central role for this commitment one's habits of inferring. In the five leading characteristics of belief cited from Tuomela, characteristic 4 postulates that beliefs are subject to an ideal of integration or agglomeration. This characteristic suggests that a reason for clinging to a belief might be its integration with other beliefs that an agent holds. In other words, consistency could be important. These characteristics could be reasons why an agent is reluctant to give up a proposition that he/she believes.

So if we try to explore the minimal characteristics of belief given in the list of ten characteristics of belief we attributed to Peirce, we have to start looking into conditions for retractions of belief, and conditions for supporting them by evidence. This takes us into matters of burden of proof, a concept related to speech acts like that of making an assertion, as opposed to saying something without making any claim that it is true and needs to be defended by giving reasons to support it.

It might be added that Peirce's observation that belief can be firmly fixed, as with fanatical believers, should be considered quite important as well. The term 'belief' is sometimes associated with fanatical believers whose belief is characterized by an unwillingness to change or retract in the face of evidence to the contrary. Belief of this sort is firmly fixed and highly resistant to rational arguments that raise doubts or present contrary evidence. The pragmatic definition of belief brings out this aspect, which might perhaps be called the irrational nature of belief, quite well.[6] We might say that belief is sometimes rational and sometimes irrational, and that therefore it would be an error in defining the notion of belief to drift into seeing it as rational belief, excluding the elements that are not influenced by rationality, or even incompatible with the requirements of rationality and evidence. This characteristic ties in with Peirce's characterization of belief as a calm and satisfactory state. Commitment can be either settled or tentative, whereas belief is a comparatively settled state in which the believer

[6] Here again it might be noted that we have stopped short of expanding the analysis to a discussion of collectively held or attributed beliefs in a group of agents, like a corporation or a religious sect.

clings to a proposition and is reluctant to give it up. Belief has a characteristic that might be called inertia. Because it is an indication of habit and is a settled state, it tends to be hard to dislodge, even, in many cases, in the presence of evidence to the contrary. This stability can be hardened, and become inflexible in some instances, as in the case of "true believers" of a dogma.

When it comes to defining a notion of belief that could be useful in artificial intelligence and multi-agent systems, the best approach is to characterize a belief as a special type of commitment that is more than just commitment for the sake of argument, and that is characterized by a stable state in an agent that is not easily changed, even though it is a matter of degree. If the agent shows persistence in clinging to a particular commitment over a sequence of deliberative or investigative reasoning, and shows reluctance to give it up, even in the face of doubts that are raised or arguments to the contrary, such indications should be taken as evidence that this commitment is a belief of the agent.

6. General Methods for Determining Belief in CBV

So how can we use ASDV to help bridge the gap between commitment and belief? Believing a proposition is not only acceptance of it, but also an internal attachment to that proposition. A belief is a proposition that an agent has embraced, so to speak, and holds personally. As such, belief is a subjective notion, a private notion, and a deeply psychological notion. What characterizes belief is a positive pro-attitude of attachment toward a proposition. To say 'Amanda believes proposition *A*' means that Amanda embraces *A* as true, in the sense that she is personally attached to *A* in a way that makes it firmly acceptable for her as a personal commitment.

ASDV is designed to model this notion of belief by providing a basic dialogue structure onto which assumptions of varying strengths can be added. The central epistemological problem about belief posed by using ASDV is how one can devise helpful methods for reasonably judging what an agent's belief is, or is not, from the data given in an instance. In other words, the question is one of how one agent should reasonably draw inferences about whether a given proposition is believed by another agent or not. We can go by what that agent has said or professed, given the evidence of how he or she has argued or acted in the past. The problem is how to infer from commitment to belief.

Hamblin's remark that commitment is a kind of *persona* of belief suggests an approach to solving this problem. This approach can be sketched out as follows. As a dialogue proceeds, a good deal of evidence may accumulate that may specify an agent's commitment concerning some proposition or issue. The agent may strongly defend a particular proposition or position in several discussions, and she may also make efforts to clarify her own attitude towards it. She might even declare that she believes this proposition, or that she does not believe it. This kind of dialogue would be evidence with respect to not just her commitment, but also her belief. The approach to solving the problem of drawing inferences from commitment to belief can thus be expressed as follows. In a given case, textual evidence of what has taken place in dialogue exchanges can sometimes take the form of a package that enables an inference from commitment to belief to be reasonably drawn. This package contains not only a consistent set of

commitments that tie in together and reinforce each other, but also textual indications that imply that the speaker is personally attached to this proposition and embraces it, as shown by strong and consistent personal declarations of commitment to it. When this kind of evidence is present, we are justified in claiming that not only is the agent committed to *A*, she also believes *A*.

What makes a belief inferable from the commitments of a dialogue participant is that it is relatively stable over the dialogue. Stability means that it tends to be (i) a commitment that is consistent with that participant's other commitments, and (ii) a commitment that the participant is reluctant to retract, even under repeated questioning. Stability means that a belief is a commitment that is closely tied in with other commitments that the participant has. For example, if a commitment is identified as a belief of a participant, then another proposition that is an immediate consequence of the proposition contained in that belief will also be highly likely to be a belief of that participant. But it is not necessarily so. Sometimes people are inconsistent in their commitments, and this can be used as a basis for questioning about a belief.

There are various methods one participant in a dialogue can use to test whether a particular proposition is a belief of the other participant. The first method is that of simply posing a question to ask the other party whether he believes this proposition or not. The second method is to examine what the other participant has gone on record as saying in the previous sequence of dialogue exchanges to see whether there is any evidence indicating that he believes this particular proposition. The third method is to examine actions that can be attributed to the other participant, and to draw inferences from these actions using as premises also the data of what the other participant has gone on record as saying in the previous sequence of dialogue exchanges. The combined use of the latter two techniques is familiar in the literature on the *ad hominem* fallacy of the circumstantial variety. In such cases, the participant may have gone on record as saying one thing but doing another, where he recommended a particular course of action as something everyone should do, but then an examination of his own personal circumstances reveals that he has failed to act in this way himself in the past, or has acted in a manner that is inconsistent with what he now recommends.

Examination dialogue is a type of dialogue that has two goals (Walton, 2006). One is to extract information to provide a body of data that can be used for argumentation in an embedded dialogue, like a persuasion dialogue for example. Examination dialogue can be classified as a species of information-seeking dialogue, and the primary goal is the extraction of information. However, there is also a secondary goal of testing the reliability of the information. Both goals are carried out by asking the respondent questions and then testing the reliability of the answers extracted from him. The formal analysis of the structure of examination dialogue by Dunne, Doutre and Bench-Capon (2005) models this testing function of examination dialogue. In their model, the proponent wins if she justifies her claim that she has found an inconsistency in the previous replies of the respondent. Otherwise the respondent wins. To implement this testing function, the information initially elicited is compared to other statements or commitments of the respondent, other known facts of the case, and known past actions of the respondent. This process of testing sometimes takes the form of attempts by the questioner to trap the respondent in an inconsistency, or even in using such a contradiction to attack the respondent's ethical character. Such a character attack used in

cross-examination of a respondent can often be used as an *ad hominem* argument, where for example, the testimony of a witness is impeached by arguing that he has lied in the past, and that therefore what he says now is not reliable as evidence.

So far we have not addressed the question of which forms of inference are used in CBV to draw inferences from a participant's commitment-set to a statement judged to be the belief of a participant on the basis of the inference. We have now mentioned two argumentation schemes that might be used for this purpose, the schemes for practical reasoning and *ad hominem* argument, in addition to *modus ponens*. The technical problem is that CBV only allows for deductive forms of inference, and does not accommodate defeasible argumentation schemes. However, many of the most important schemes that would be most useful for this purpose, the scheme for argument from commitment for example, are defeasible. Indeed we have argued of that an advantage of the methodology for belief posed above is that it uses the device of critical questions matching an argumentation scheme. In the next section we discuss how CBV can be extended to a system that admits of argumentation schemes of this sort.

7. Determining an Agent's Beliefs

The traditional problem of other minds in philosophy is that one person has no direct way of determining what another agent's internal states (like beliefs) are. This problem, curiously, is also present in artificial intelligence. Designers of systems of software agents in multi-agent computing are generally not able to access the internal beliefs, motivations or goals of the agents participating in their system, because many of the agents will be designed and created by different design teams. Even if these internal states were observable to a system designer, a sufficiently-clever agent designer can always produce an agent able to pass any test of its internal states imposed by the system designer. This situation is called problem of semantic verification in multi-agent systems (see Wooldridge 2000a). It has led researchers in multi-agent systems to concentrate on assessing an agent's states only through externally observable behaviors. The same problem recurs in trying to understand animal behavior where, as we noted in section 1, researchers on chimp behavior used the notion of simulation.

The notion of simulation offers a clue into how one agent typically goes about drawing conclusions about what another agent's commitments and beliefs are. The two agents share forms of reasoning, like practical reasoning. The one agent's words and deeds can used as data furnishing premises from which the other agent can draw inferences. These forms of reasoning, it can be argued, include deductive and inductive forms, as well as presumptive schemes, like argument from expert opinion, that represent a third type of argumentation, plausible reasoning. Explicit commitments are public and social, but implicit commitments are, in many instances, not directly stated by an agent, and therefore have to be inferred indirectly, based on unstated premises. But it is fairly clear how commitments should generally be determined, given the evidence of an agent's words and deeds in a case. This task requires examining not only what he actually said, or went on record as saying, but looking beneath this to try to fairly determine what this commits him to by drawing inferences from what he said.

Beliefs are derived abductively by one participant in a dialogue from the commitment set of the other participant using evidence collected so far in the dialogue. The theory is

based on the notion of simulation, whereby one agent, called the secondary agent, draws an inference in the form of a hypothesis about another agent's belief. The secondary agent draws the hypothesis by looking at the evidence of the primary agent's words and deeds, combined with what it knows about the agent's commitments. In the case of autoepistemic reasoning the agent can form a hypothesis about what her/his own belief is. This method is based on the ASDV system, according to which an agent's belief is a specially designated proposition that is selected from his/her commitments. On this theory, a belief is a special type of commitment that has remained stable throughout the course of a deliberation or investigation, and that fits in and agglomerates with other propositions known to be beliefs of that agent. Thus there are three components involved in identifying a belief. The first component is that there needs to be a distinction between the primary agent who possesses the belief and the secondary agent judging that that other agent has that belief or not.[7] Each plays a different role in the process of determining whether a given proposition represents a belief of the primary agent or not. So you could say that there are always two participants in a dialogue, where each agent has a role as a participant in an investigation or deliberation. The second component is the evidence on which the hypothesis identifying a proposition as a belief is based. The third component is the inference by which the hypothesis is inferred from the evidence. An especially important type of inference is that of abductive reasoning, or inference to the best explanation, which proceeds by selecting a best explanation from a set of data.

The basic argumentation scheme for abductive reasoning (Walton, Reed and Macagno, 2008, 329) is presented below.

PREMISE 1: D is a set of data or supposed facts in a case.
PREMISE 2: Each one of a set of accounts $A_1, A_2, ..., A_n$ is successful in explaining D.
PREMISE 3: A_i is the account that explains D most successfully
CONCLUSION: Therefore A_i is the most plausible hypothesis in the case.

Matching this scheme is a set of critical questions (Walton, Reed and Macagno, 330) that one party in a dialogue can use to raise doubts concerning whether the application of the scheme to a particular case is justified.

CQ_1: How satisfactory is A_i itself as an explanation of D, apart from the alternative explanations available so far in the dialogue?
CQ_2: How much better an explanation is A_i than the alternative explanation so far in the dialogue?
CQ_3: How far has the dialogue progressed? If the dialogue is an inquiry, how thorough has the search been in the investigation of the case?
CQ_4: Would it be better to continue the dialogue further, instead of drawing a conclusion at this point?

Abductive reasoning is based on the factual evidence given in a case at issue. The kinds of evidence used to draw a hypothesis about an agent's beliefs come from the data

[7] In autoepistemic reasoning, as noted above, these agents are the same.

possessed in the case are based on how the agent has acted, and the arguments and other speech acts the agent put forward during the course of the deliberation or investigation.

There are two kinds of evidence, positive and negative. Positive evidence, of the following eight types, is evidence supporting the hypothesis that constitutes the explanation of the data. (1)The first kind of positive evidence is explicit or implicit commitment to the proposition in question. (2) If the agent has made strong and emphatic assertions that a particular proposition is true, this would be evidence that he/she believes this proposition. (3) Another kind of evidence is that if under varying circumstances the agent clings tenaciously to affirming, defending, or acting on that proposition, or behaving in a way that indicates it accepts it as true, it is a belief of that agent. (4) Resistance to attempts to refute the proposition is another kind of evidence. (5) Responding to requests for justification by giving evidence for the proposition, rather than retracting it, is another kind of positive evidence. (6) Another kind of evidence is the stability in the way the agent continues to maintain and defend this proposition, especially where that evidence indicates that no doubt is being expressed about whether the proposition is true. (7) Another kind of evidence concerns indications of a habit of defending this belief, accepting it, advocating it, or acting on it. (8) Another kind of evidence concerns the practical reasoning of the agent as it engages in goal-directed actions. If a particular proposition is closely connected to actions used by an agent as means to carry out goals, or closely connected to the goals themselves, that is evidence that the proposition in question is a belief, and not just a commitment of that agent.

Negative evidence concerns alternative explanations that need to be considered, and ruled out in order to establish something as a belief. It includes instances where an agent professes to believe a proposition, but there is also evidence of insincerity, self-deception, hypocrisy or lying. Evidence of insincerity, or even hypocrisy, can be found in cases where a person claims to accept a particular proposition, but acts in a way that is inconsistent with accepting it. Hypocrisy is closely related to inconsistency of commitments, and based on it, but inconsistency is in some instances itself a type of negative evidence that goes against the hypothesis that an agent holds a belief.

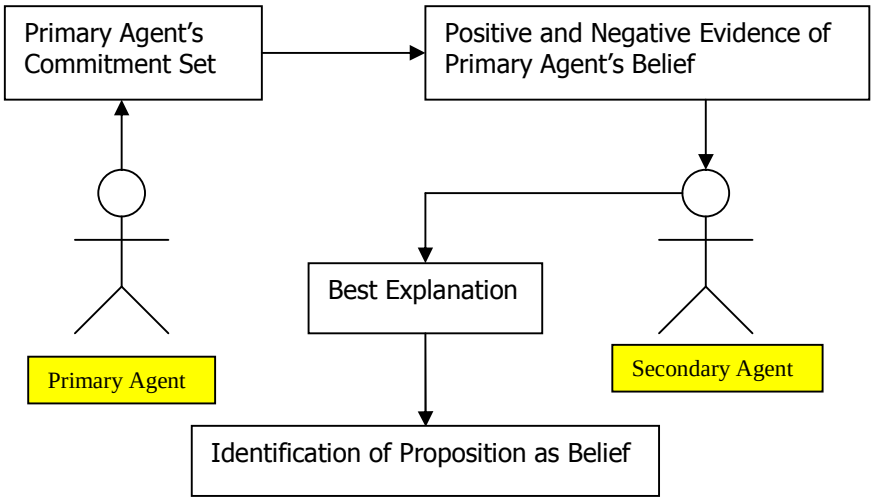

Figure 1: Using Abduction to Derive an Agent's Belief by Inference

The procedure whereby this evidence is employed to draw a conclusion about an agent's belief by abductive reasoning is shown graphically in figure 1. As shown in figure 1, the secondary agent begins with the commitment set of the primary agent, and from the evidence of what the primary agent said and did as known in the circumstances of the case, draws a hypothesis on whether a particular commitment attributed to the primary agent can properly be said to be one of its beliefs.

8. Conclusions

This section fits the definition of belief previously formulated in the paper into a more precise formulation in the structure of ASDV and presents an argumentation scheme for arguments from commitment to belief. First, we need to recall the characteristics of belief that emerged from the discussion leading up to the definition. It was emphasized that (1) belief is a settled state (2) that we cling to, (3) sometimes very strongly, (4) it is opposed to doubt, an uneasy or unsettled state, (5) it is shaped by evidence, and (6) it guides our actions and argumentation. It was also stated that (7) belief is a matter of degree. The discussion of these characteristics then led us to a definition of 'belief' that can be summarized as follows: belief is a stable state of an agent that is not easily changed, is a matter of degree (held more or less weakly or strongly), and that guides the desires and actions of that agent in goal-directed reasoning in deliberation, or the evidential thinking of that agent in an investigation aimed at removing doubt. Some other characteristics were mentioned as well, like the agglomeration of beliefs into groups.

In the paper, the main thesis argued for was that beliefs can be inferred from commitments, but not in a straightforward manner. The thesis was that we start with the set of an agent's commitments in a dialogue, pick out some of them as potential beliefs of that agent, and then run them through the list of seven characteristics cited above to determine whether they meet the requirements for them. If so, they can be hypothesized to be beliefs of the agent, subject to further evidence that might come in as the dialogue proceeds through the argumentation stage. This analysis is the basis for formulating the following basic defeasible argumentation scheme for argument from commitment to belief in, where *a* is an agent taking part in a dialogue *D* built on the basic structure of ASDV.

PREMISE 1: *a* is committed to *A* in a dialogue *D* based on an explanation of *a*'s commitments in *D* in the dialogue.
PREMISE 2: *a*'s commitment to *A* is not easily retracted under critical questioning in *D*.
PREMISE 3: *a*'s commitment to *A* is used as a premise in *a*'s practical reasoning and argumentation in *D*.
CONCLUSION: Therefore *a* believes *A* (more strongly or weakly).

This scheme is built on the assumption that there is some way of ordering the comparative weakness or strength of the propositions in an agent's set of beliefs, representing how firmly the agent is committed to that belief. Such firmness is indicated

by how easily the proposition is given up under critical questioning by the other party in the dialogue, and by how prominently it is used as a premise in *a*'s argumentation.

The following critical questions match the basic scheme for argument from belief commitment.

CQ_1: What evidence can *a* give that supports his belief that *A* is true?
CQ_2: Is *A* consistent with *a*'s other commitments in the dialogue?
CQ_3: How easily is *a*'s commitment to *A* retracted under critical questioning?
CQ_4: Can *a* give evidence to support *A* when asked for it?
CQ_4: Is there some alternative explanation of *a*'s commitments?

Included in *a*'s commitments are *a*'s goals, actions and professed beliefs.

It is also useful to have a comparative scheme for argument from commitment to belief with the conclusion that *a* believes more strongly that *A* than that *B*.

PREMISE 1: *a* is committed to *A* more strongly than *B* in a dialogue *D* based on *a*'s explicit or implicit commitments in *D* in the sequence of dialogue.
PREMISE 2: *a*'s commitment to *A* is less easily retracted under critical questioning in *D* than *a*'s commitment to *B*.
PREMISE 3: *a*'s commitment to *A* is used as a premise in *a*'s practical reasoning and argumentation in *D* more often and centrally than *a*'s commitment to *B*.
CONCLUSION: Therefore *a* believes *A* more strongly than *B*.

The critical questions matching the comparative scheme are the following.

CQ_1: How stable is *a*'s commitment to *A* over *B* during the course of *D*?
CQ_2: Is there evidence from the alternative explanations available so far in *D* suggesting that *a* does not believe *A* more strongly than *B*?
CQ_3: How easily is *a*'s tenacity of commitment to *A* rather than to *B* retracted under critical questioning?
CQ_4: Can *a* give stronger evidence to support *A* when asked for it rather than to the evidence he gives to support *B* when asked for it?

Both sets of critical questions are meant to fit with the five widely accepted properties of belief and the ten defining properties of belief attributed to Peirce in section 3.

This account of belief uses defeasible argumentation schemes that correspond to the habits of inferring that Peirce took to be fundamental to the notion of belief. Both schemes and their matching critical questions are based on Peirce's view that beliefs are subject to an ideal of integration or agglomeration. They reflect the list of the five widely accepted characteristics of belief set out in section 3 by incorporating the assumptions that beliefs aim at truth, and that they are shaped by evidence for what is believed. Much more needs to be said about how belief works in different types of dialogues, for example in examination dialogue and deliberation dialogue. These are left as projects for future research.

As indicated above, finding an agent's actual belief can be very difficult in some instances, even after extensive interrogation and argumentation, and finding conclusive

proof that an agent believes a particular proposition is typically not possible. In many instances, there is not enough evidence to say with a high degree of confidence that an agent believes a particular proposition. Often it is very hard for us to even know what our own beliefs are, even after considerable introspection and discussion. In this typical kind of situation of lack of direct accessibility to the external evidence, basing the method of determining belief on abductive reasoning is highly suitable, as is the dialogue method. The dialogue method is aided by work on one special type of dialogue called examination (Dunne et al., 2005; Walton, 2006). However, to give a general model of belief that can be carried further to many more specialized of dialogue like examination, we used a basic system CBV and extended it to ASDV so that various argumentation schemes, especially the ones for practical reasoning and abduction, can be used as rules of inference. The general theory is that each participant in a dialogue can use such schemes to critically question each other's statements and arguments by drawing inferences from each others' commitment sets and then use them to draw out evidence-based conjectures about the beliefs of the other party.

References

Katie Atkinson, Trevor Bench-Capon and Peter McBurney, 'Computational Representation of Practical Argument', *Synthese*, 152, 2006, 157-206.

Trevor J. M. Bench-Capon, 'Persuasion in Practical Argument Using Value-based Argumentation Frameworks', *Journal of Logic and Computation*, 13, 2003, 429-448.

Michael Bratman, *Intention, Plans and Practical Reason*, Cambridge, Mass., Harvard University Press, 1987.

Michael Bratman, 'Practical Reasoning and Acceptance in a Context', *Mind*, 1-2, 1993, 1-15.

Michael E. Bratman, David J. Israel and Martha E. Pollack, 'Plans and Resource-bounded Practical Reasoning', *Computational Intelligence*, 4, 1988, 349-355.

Paul Churchland, 'Eliminative Materialism and the Propositional Attitudes', *Journal of Philosophy*, 78, 1981, 67-90.

Jonathan Cohen, *An Essay on Belief and Acceptance*, Oxford, Oxford University Press, 1992.

Paul E. Dunne, Silvie Doutre and Trevor J. M. Bench-Capon, Discovering Inconsistency through Examination Dialogues, *Proceedings IJCAI-05 (International Joint Conferences on Artificial Intelligence)*, Edinburgh, 2005, 1560-1561. Available at: http://ijcai.org/search.php

Paul E. Dunne and Trevor J. M. Bench-Capon, eds, *Computational Models of Argument: Proceedings of COMMA 2006*, Amsterdam, IOS Press, 2006, 195-207.

Pascal Engel, 'Believing, Holding True, and Accepting', *Philosophical Explorations*, 1, 1998, 140-151.

Pascal Engel (ed.), *Believing and Accepting*, Dordrecht, Kluwer, 2000.

David M. Godden, 'The Importance of Belief in Argumentation', *Synthese*, to appear, 2009.

Charles L. Hamblin, *Fallacies*, London, Methuen, 1970.

Charles L Hamblin, C. L., 'Mathematical Models of Dialogue', *Theoria*, 37, 1971, 130-155.

Wolfram Hinzen, Review of Engel (2000), *Grazer Philosophische Studien*, 62, 2001, 282, 286.

Mark Kaplan, 'Rational Acceptance, *Philosophical Studies*, 40, 1981, 129-146.

Isaac Levi, *Rational Acceptance*, Cambridge, Cambridge University Press, 1997.

Nicholas Maudet and Brahim Chaib-draa, 'Commitment-based and Dialogue-Game Based Protocols: New Trends in Agent Communication Languages', *The Knowledge Engineering Review*, 17, 2002, 157-179.

Anthonie Meijers, 'Collective Agents and Cognitive Attitudes', *Protosociology*, 16, 2002, 70-85.

Fabio Paglieri and Cristiano Castelfranchi, 'Arguments as Belief Structures', *The Uses of Argument: Proceedings of a Conference at McMaster University*, ed. David Hitchcock and Daniel Farr, Ontario Society for the Study of Argumentation, 2005, 356-367.

Charles S. Peirce, 'The Fixation of Belief', *Popular Science Monthly* 12 (November 1877), 1-15.

Chris Reed and Douglas Walton, 'Argumentation Schemes in Dialogue', *Dissensus & the Search for Common Ground: Proceedings of OSSA*, June 2007, Windsor, Ontario, CD-ROM, 2007, 1-11: http://www.dougwalton.ca/papers%20in%20pdf/07OSSAchris.pdf

Lynne Rudder Baker, *Saving Belief*, Princeton, Princeton University Press, 1989.

John R. Searle, *Rationality in Action*, MIT Press, Cambridge, MA, USA, 2001.

Munindar P. Singh, 'Agent Communication Languages: Rethinking the Principles', *Computer*, 31, 1998, 425-445.

Stephen Stitch, *From Folk Psychology to Cognitive Science*, Cambridge, Mass., The MIT Press, 1983.

Raimo Tuomela, 'Belief Versus Acceptance', *Philosophical Explorations*, 2, 2000, 122-137.

Frans H. van Eemeren and Rob Grootendorst, *Argumentation, Communication and Fallacies*, Hillsdale, N.J., Lawrence Erlbaum Associates, 1992.

Douglas Walton, *Logical Dialogue-Games and Fallacies*, Lanham, Maryland, University Press of America, 1984. Available at http://www.dougwalton.ca/books/LDG84bk.pdf

Douglas Walton, *Practical Reasoning: Goal-Driven, Knowledge-Based, Action-Guiding Argumentation*, Savage, Maryland, Rowman & Littlefield, 1990.

Douglas Walton, 'Examination Dialogue: An Argumentation Framework for Critically Questioning an Expert Opinion', *Journal of Pragmatics*, 38, 2006, 745-777.

Douglas Walton and Erik C. W. Krabbe, *Commitment in Dialogue*, Albany, State University of New York Press, 1995.

Douglas Walton, Chris Reed and Fabrizio Macagno, *Argumentation Schemes*, Cambridge, Cambridge University Press, 2008.

Michael Wooldridge, *Reasoning about Rational Agents*, Cambridge, Mass., The MIT Press, 2000.

Michael Wooldridge, 'Semantic Issues in the Verification of Agent Communication Languages', *Journal of Autonomous Agents and Multi-Agent Systems*, 3, 2000a, 9-31.

Michael Wooldridge, *MultiAgent Systems*, Chichester, Wiley, 2002.